На правах рукописи

Гильмуллин Ринат Абрекович

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В МНОГОЯЗЫКОВЫХ СИСТЕМАХ ОБРАБОТКИ ДАННЫХ НА ОСНОВЕ АВТОМАТОВ КОНЕЧНЫХ СОСТОЯНИЙ

05.13.11 - Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Казань – 2009

Работа выполнена на кафедре теоретической кибернетики государственного образовательного учреждения высшего профессионального образования «Казанский государственный университет им. В.И. Ульянова-Ленина»

Научный руководитель: академик АН РТ,
доктор технических наук, профессор
Сулейманов Джавдет Шевкетович

Научный консультант: доктор физико-математических наук,
доктор технических наук, профессор
Бухараев Раис Гатич

Официальные оппоненты: доктор физико-математических наук,
профессор
Елизаров Александр Михайлович

доктор технических наук, профессор
Соснин Пётр Иванович

Ведущая организация: Московский государственный университет, НИВЦ, г. Москва

Защита состоится «21» января 2010 г. в 16:00 на заседании диссертационного совета Д 212.081.24 при Казанском государственном университете им. В.И. Ульянова-Ленина по адресу: 420008, г. Казань, ул. Кремлевская, д. 18, конференц-зал научной библиотеки им. Н.И. Лобачевского.

С диссертацией можно ознакомиться в научной библиотеке им. Н.И. Лобачевского Казанского государственного университета.
Автореферат разослан «18» декабря 2009 г.

Учёный секретарь
диссертационного совета,
к. ф.-м. н., доцент Еникеев А.И.

Общая характеристика диссертации

Актуальность проблемы. В системах обработки естественно-языковых (ЕЯ) текстов, таких как системы машинного перевода, системы автоматизированной коррекции текстов, системы многоязыкового поиска в локальных базах данных и сети Интернет, значительное место занимает процесс математического моделирования лингвистических структур для эффективной целевой обработки данных. Существенные результаты в этих областях получены в работах российских и зарубежных исследователей Д.А. Поспелова, И.А. Мельчука, В.Ф. Хорошевского, Г.С. Осипова, Ю.Д. Апресяна, И.М. Богуславского, Л.Л. Цинмана, Л.Л. Иомдина, А.С. Нариньяни, М.Г. Мальковского, Б.В. Доброва, Н.В. Лукашевич, Т.А. Гавриловой, Р.Г. Бухараева, Д.Ш. Сулейманова, П.И. Соснина, О.А. Невзоровой, С.А. Шарова, Ю.Р. Валькмана, Н. Хомского, Р. Каплана, М. Кея, К. Коскенниеми и др.

Математическое моделирование лингвистических структур (разработка математических лингвистических моделей) – это, по сути, научно-прикладная область фундаментальных исследований для анализа, синтеза, интерпретации и трансформации ЕЯ текстов[1]. Построение систем обработки данных (СОД) на основе универсальных лингвистических моделей практически невозможно ввиду отсутствия универсальной, или даже достаточно полной формальной модели какого-либо языка, и сложности вычислительной реализации универсальных СОД (в общем случае задача является NP полной).

Одним из способов повышения эффективности построения СОД является концепция прагматически-ориентированного подхода к разработке математических лингвистических моделей, определяющий минимальный набор средств для решения определенного круга лингвистических задач, исходя их принципа достаточности[2].

Прагматически-ориентированный подход к построению лингвистических моделей это, прежде всего, концептуально-инструментальная технология, которая позволяет, с одной стороны, осуществлять адекватный подбор средств эффективной обработки ЕЯ-текста, с другой стороны, детерминировать контекст и управлять формированием образа генерируемого или распознаваемого ЕЯ-текста.

В частности, вычислительная сложность разработки алгоритмов может быть снижена за счет учета специфики языковых данных, уровней детализации и глубины разработки математических лингвистических моделей различных языковых уровней. Предметом исследования в диссертации являются математические лингвистические модели родственных языков (на примере тюркских языков), которые характеризуются общим набором параметров описания на

---

[1] *Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др.* Лингвистический процессор для сложных информационных систем. М.: Наука, 1992.

[2] *Сулейманов Д.Ш.* Системы и информационные технологии обработки естественно-языковых текстов на основе прагматически-ориентированных лингвистических моделей. Диссертация на соискание ученой степени доктора технических наук. 2000.

всех языковых уровнях. Параметры морфологической модели родственных языков во многом определяют параметры синтактико-семантической модели предложения. Следовательно, актуальной и перспективной является задача разработки математических лингвистических моделей и базовых программных технологий обработки текстов для многоязыковых систем обработки данных одной языковой группы.

Цель и задачи исследования. Целью диссертационной работы является исследование, разработка и реализация математических лингвистических моделей и программного обеспечения систем и технологий обработки многоязыковой информации.

Для достижения поставленной цели в рамках диссертационной работы решаются следующие основные задачи:

- Исследование и разработка автоматной модели лингвистических формализмов на примере татарской морфологии;
- Реализация программных модулей генерации и распознавания морфологии тюркских языков;
- Разработка формальных семантических моделей аффиксальных морфем на основе объектно-предикативных схем и проведение сопоставительного анализа семантических схем для тюркских языков;
- Разработка формальной модели перевода на основе алгоритмов машинного обучения, использующих шаблоны переводных соответствий тюркских языков;
- Реализация программных модулей системы машинного перевода тюркских языков.

О**бъект исследования.** Объектами исследования являются:

1) Двухуровневая автоматная модель морфологии тюркских языков;
2) Объектно-предикативные схемы для формальных семантических моделей аффиксальных морфем тюркских языков;
3) Формальная модель перевода на основе алгоритмов машинного обучения.

Как отмечалось выше, предметом исследования являются математические лингвистические модели родственных языков на примере татарского и турецкого языков. Выбор этих языков обусловлен их общими типологическими характеристиками, в частности, общей регулярной морфологией, а также общими структурно-функциональными моделями предложений, что является существенным для перевода.

Научная новизна результатов. В процессе исследований получены следующие новые научные результаты, выносимые на защиту.

Полная компьютерная модель татарской морфологии в виде двухуровневой автоматной модели.

Программный инструментарий для морфологического анализа и синтеза татарских текстов на основе двухуровневой автоматной модели морфологии.

Формальные семантические модели аффиксальных морфем на основе объектно-предикативных схем.

Формальные модели перевода на основе алгоритмов машинного обучения, использующие шаблоны переводных соответствий тюркских языков.

Программные модули в составе системы татарско-турецкого машинного перевода.

Работа имеет принципиальную новизну, как в постановке задачи, так и в выборе методов решения поставленной задачи. Эффективность методов и подходов решения поставленной проблемы базируется, прежде всего, на комплексном использовании современных достижений в области искусственного интеллекта, математической лингвистики и компьютерных технологий, связанных с разработкой формальных моделей языка, теории и практики машинного перевода.

Практическая ценность полученных результатов.

Полученные результаты (разработанные математические лингвистические модели) успешно используются в учебном процессе в Казанском государственном университете и в Татарском государственном гуманитарно-педагогическом университете в учебных курсах “Представление и обработка знаний”, “Математическая лингвистика” и др., в научных исследованиях, проводимых на факультете татарской филологии и истории КГУ и Института языка, литературы, искусства АНТ им. Г. Ибрагимова, а также в мультимедийных учебных разработках НИИ «Прикладная семиотика» Академии наук РТ и Казанского государственного университета. Разработанные программные модули татарской морфологии внедрены в состав системы оптического распознавания текстов FineReader компании ABBYY, а также в состав Университетской информационной системы РОССИЯ (НИВЦ МГУ) для поддержки многоязычного поиска в татарско-русской электронной коллекции текстов. Разработанная формальная модель турецкой морфологии используется в многоязычном электронном словаре ABBYY Lingvo x3.

Предложенная двухуровневая автоматная модель морфологии может быть использована в составе специализированных систем, таких как автоматизированное рабочее место лингвиста.

Одной из главных особенностей построенных систем, обеспечивающих ее эффективность и гибкость, является разделенное представление языконезависимых и языкозависимых блоков. Это позволяет легко модифицировать лингвистическую базу системы, а также наполнять ее лингвистическими ресурсами, правилами, лингвистическими моделями другого языка, а также модифицировать программные модули без изменения лингвистических ресурсов.

Практические разработки и реализация результатов диссертации осуществлялись в рамках Государственной программы Республики Татарстан по сохранению, изучению и развитию языков народов Республики Татарстан.

Документы, подтверждающие внедрение и практическое использование результатов диссертации, прилагаются.

Методы исследования. При разработке и реализации двухуровневой автоматной модели морфологии использовались теория формальных грамматик и теория конечных автоматов.

Методы структурного и сопоставительного анализа, когнитивного моделирования и математической лингвистики применены при описании объектно-предикативных схем, используемых для перевода.

При разработке математических лингвистических моделей и программных модулей обработки многоязыковых данных использовались современные методы и технологии программирования.

Апробация работы. Результаты работ докладывались автором на международных конференциях и семинарах: на Международной конференции LP'2000 по типологии языков (Чехия, г.Прага, 2000), на научном семинаре по ЕЯ-процессорам в Белкентском университете (Турция, г.Анкара, 1997), на Международной конференции "KDS" (Крым, г.Ялта, 1997), на международных семинарах по компьютерной лингвистике и ее приложениям ДИАЛОГ (г.Таруса, 1998, 1999, г.Протвино, 2000-2003), на Международной конференции "Языковая семантика и образ мира" (г.Казань, 1997), на Международной конференции "Интерактивные системы: проблемы человеко-компьютерного взаимодействия" (г.Ульяновск, 2001, 2009), на Казанской школе-семинаре по компьютерной и когнитивной лингвистике TEL (г.Казань, 1999-2009), на Международном симпозиуме "LENCA-2" (г.Казань, 2004), на Международном симпозиуме «Языковые контакты Поволжья» (г. Казань, 2008), на телеконференции "Информационные технологии в гуманитарных науках" (КГУ, 1998), а также на различных республиканских и городских научных семинарах, итоговых научных конференциях КГУ и ИЯЛИ АНТ (1997-2009).

При непосредственном участии автора выполнено шесть научно-исследовательских грантов: 1) Грант Программы "Наука за стабильность" в рамках проекта TU-Language: "Татарский двухуровневый морфологический анализатор" (1996-1998 гг.). 2) Грант НИОКР АН РТ. "Разработка татарско-русского машинного переводчика регистрационных форм" (2001-2003 гг.). 3) Грант НИОКР АН РТ. "Компакт-диск с татарской локализацией об Академии наук Татарстана к 10-летнему юбилею АНТ" (2000-2001 гг.). 4) Грант НИОКР АН РТ "Машинный фонд татарского языка" (2002-2004 гг.). 5) Грант РФФИ (№ 04-06-97501) "Прикладная грамматическая модель татарского языка в задачах информационного поиска в многоязычных корпусах текстов" (2006 г.). 6) Грант РФИИ (№04-06-97501) "Экспериментальная загрузка многоязычной (русско-татарской) текстовой коллекции и адаптация соответствующих программных интерфейсов к татарскому языку на базе программных средств Университетской информационной системы "УИС РОССИЯ" (2007-2008 гг.).

За цикл работ по темам «Построение базовых программных модулей системы татарско-турецкого машинного перевода» и «Татарская локализация операционной системы Windows Vista и пакета Microsoft Office-2007» в 2004 и в 2008 годах соответственно Указом Президента Республики Татарстан и Постановлением Кабинета Министров Республики Татарстан результаты диссертации удостоены Республиканской премии молодых ученых в области «Информатика, вычислительная техника и автоматизация».

Инновационный проект «Татсофт 3: информационно-программный комплекс поддержки татарского языка в инфо-коммуникационных технологиях», включающий результаты исследований и разработок диссертанта, стал победителем на Республиканском конкурсе инвестиционно-венчурного фонда «50 лучших инновационных идей 2007 года для Республики Татарстан».

Основные результаты, полученные соискателем в рамках диссертационной работы, вошли в состав научно-образовательной темы «Научное, учебно-методическое и информационно-программное обеспечение реализации татарского языка как государственного в системе образования Республики Татарстан», удостоенной Государственной премии Республики Татарстан в области науки и техники за 2009 год.

Структура и объем работы. Диссертация состоит из введения, трех глав, заключения, списка использованной литературы и шести приложений. Объем диссертации составляет 150 страниц, 20 таблиц, 15 рисунков.

Краткое содержание диссертации

Во **введении** обоснована актуальность темы, сформулирована цель работы и определен перечень решаемых задач, указана их новизна, отмечены особенности подхода, раскрываемого в диссертационной работе, теоретическая и практическая ценность полученных решений и разработок, а также дан краткий обзор содержания по главам.

В первой главе дается аналитический обзор разработок и литературы по теме диссертации.

Проведен анализ систем и методов в области систем машинного перевода, который позволил сформулировать подход к разработке концепции и методологии программно-концептуальной прагматически-ориентированной технологии для создания систем машинного перевода тюркских языков.

На основе анализа формальных моделей и средств обработки ЕЯ-текстов сделан вывод о том, что двухуровневая автоматная модель морфологии, являющаяся прагматически-ориентированной формальной моделью, может быть эффективно использована при разработке систем машинного перевода для тюркских языков.

Раскрывается постановка основных задач диссертации.

Во второй главе описывается математическая лингвистическая модель морфологии на основе двухуровневого формализма и программная реализация моделей в составе двухуровневого морфологического анализатора.

Двухуровневый формализм представляется в нотации двухуровневых правил, которые устанавливают законы соответствия между поверхностным и глубинным уровнями представления символов в зависимости от контекста реализации.

1. Представление двухуровневых правил конечными автоматами.

Основным механизмом представления двухуровневых правил в виде двухуровневой компьютерной модели является технология автоматов конечных состояний (АКС) в виде трансдьюсеров конечных состояний (ТКС). ТКС отличается от АКС тем, что оперирует над двумя входными последовательно-

стями. Он распознает, действительно ли две последовательности являются соответствиями (т.е. переводами друг в друга).

Пример 1. Предположим, что первая входная строка для ТКС является цепочки языка *L1***,** содержащего элементы *x* и *y* и определенного как $L1=\{xy^n x | n \geq 0\}$. Правильно построенными цепочками для этого языка будут: *xx*, *xyx*, *xyyx*, *xyyyx*, и т.д. В качестве второй входной строки определим цепочки языка *L2,* соответствующие цепочкам языка *L1*, в которых каждое второе вхождение элемента *y* соответствует элементу *z*.

На рисунке 1 показана диаграмма ТКС для примера 1. Единственное отличие диаграммы ТКС от диаграммы АКС заключается в том, что дуги помечены парами соответствий, содержащих символы обоих входных языков.

ТКС также могут быть представлены в виде таблиц конечных состояний,

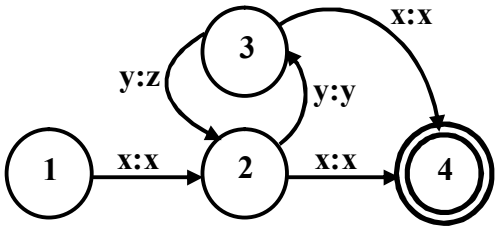

Рис. 1. Диаграмма ТКС соответствия между языками L1 и L2.

с той лишь разницей, что в заголовках столбцов будут указаны пары соответствий, такие как: *x:x, y:y* и *y:z*. Например, диаграмма, указанная на рисунке 1 может быть представлена в виде следующей таблицы конечных состояний:

	x x	y y	y z
1.	2	0	0
2.	4	3	0
3.	4	0	2
4:	0	0	0

1.1. Конструкция двухуровневых правил:

RULE L_f:S_f F C_l_C_r

Правило **RULE** состоит из трех частей:

1) **L_f:S_f – связь-соответствие,** где **L_f –** это лексический символ, **S_f** – поверхностный символ;

2) **F – оператор (функция переходов в ТКС)**, определяющий соответствие в зависимости от контекста. Имеется четыре типа оператора: **=>, <=, <=>** и **/<=**

F1) => означает, что соответствие проявляется только в этом окружении, но не всегда;

F2) <= означает, что соответствие в этом окружении проявляется всегда, но не только в этом окружении;

F3) <=> означает, что соответствие проявляется в этом окружении всегда и только в этом окружении;

F4) /<= означает, что соответствие никогда не проявляется в этом окружении.

3) $\mathbf{C_l\_C_r}$ **– контекст**, в котором допускается входное соответствие, где $\mathbf{C_l}$ обозначает левый контекст, а $\mathbf{C_r}$ – правый контекст относительно входного соответствия.

При двухуровневом подходе фонология определяется как связь между лексическим уровнем глубинного представления слов и их реализации на поверхностном уровне.

2. Файл фонологических правил создается пользователем для описания алфавита языка и фонологических правил. Файл состоит из списка объявлений ключевых слов и соответствующего им содержания. В описании файла фонологических правил используются следующие элементы формализации:

ALPHABET – список символов, необходимых для полного представления алфавита того или иного ЕЯ.

NULL 0 – фонологический процесс, который удаляет или вставляет символы в двухуровневую модель соответствующие символу **NULL**; записываются как **0** (ноль).

ANY @ – обозначает любой символ из списка **ALPHABET**.

BOUNDARY # – граничный символ**.** Обозначает границу слова – либо начало, либо конец.

SUBSET – используется для обозначения определенного множества символов.

RULE – стандартный идентификатор для двухуровневого правила.

END – признак конца файла фонологических правил.

Для описания файла фонологических правил татарского языка используется 47 правил, подробное описание которых приводится в разделах диссертации.

Ниже приведены примеры двухуровневых правил **П1** и **П2.**

П1 – двухуровневое правило, описывающее морфофонемический процесс для сонорных звуков татарского языка:

(П1) RULE " Л:н <=> SONOR +:0 _" 3 5

	Л н	Л @	Sonor Sonor	+ 0	@ @
1:	0	1	2	1	1
2:	0	1	2	3	1
3:	1	0	2	1	1

Правило **П1** означает, что лексический символ ***Л*** соответствует поверхностному символу ***н*** тогда и только тогда, когда ему предшествует сонорные согласные из множества ***SONOR (н, м, ң)***, определенные в разделе описания множеств файла фонологических правил. Благодаря этому правилу выводимы

следующие поверхностные формы: *сан+**Л**Ар ->сан<u>нн</u>ар, урам+**Л**Ар -> ура<u>мн</u>ар, таң+**Л**Ар ->та<u>ңн</u>ар.*

П2 – описывает установление соответствия лексических символов ***к*** и ***п*** поверхностным символам ***г*** и ***б*** соответственно.

(П2) RULE {к, п}:{г, б} <=> @:VOWEL_ +:0 (C:0) @:VOWEL|VOWEL_ +:0 [р [А:а|А:ә] к | [Ы:ы|Ы:е]];

Правило **П2** состоит из двух контекстов:

(а) [VOWEL|@:VOWEL]_ +:0 (C:0) @:VOWEL

(б) VOWEL_ +:0 [р [А:а|А:ә] к | [Ы:ы|Ы:е]]

Контекст (а) правила **П2** утверждает, что лексический символ ***к*** соответствует поверхностному символу ***г*** и лексический символ ***п*** соответствует поверхостному символу ***б***, если:

1) слева направо им предшествует любой лексический символ, соответствующий любому поверхностному символу из множества гласных букв **VOWEL**.

2) справа налево от них следует символ +, соответствующий ***0***, после которого может встретиться любой символ из множества ***C***, также соответствующий символу ***0***, далее следует любой лексический символ, соответствующий любому поверхностному символу из множества **VOWEL**.

Контекст (б) правила **П2** утверждает, что лексический символ ***к*** соответствует поверхностному символу ***г*** и лексический символ ***п*** соответствует поверхостному символу ***б***, если:

1) слева направо им предшествует любой лексический символ из множества **VOWEL**.

2) справа налево от них следует символ +, соответствующий ***0***, далее символ ***р***, далее лексический символ ***А***, соответствующий любым поверхностным символам ***а*** или ***ә***, далее символ ***к***, за которым следует лексический символ ***Ы***, соответствующий поверхностным символам ***ы*** или ***е***. По данному правилу выводимы следующие поверхностные формы: *кита**п**+Ым -> кита**б**ым, кала**к**+Ым -> калағым, өстә+Ы**п**+рАк -> өстә**б**рәк, а**к**+рАк -> аққрак.*

<u>3. Файл морфотактических правил</u> также является пользовательским файлом, который содержит список лексических единиц, и описание морфотактических правил. Лексическая единица может быть одной единственной морфемой (такой как корень, префикс или суффикс) или морфологическим комплексом слов (корень плюс префикс и суффикс). При распознавании слов лексические компоненты работают совместно с компонентами правил. Генеральной структурой лексикона является список объявлений ключевых слов. Множество действительных ключевых слов включает **ALTERNATION, LEXICON, INCLUDE** и **END**. Объявления могут встречаться в любом порядке за исключением того, что **LEXICON** должен объявляться после **ALTERNATION**. Обязательное единственное объявление - это **LEXICON INITIAL**; то есть, лексический файл как минимум должен содержать подлексикон, называемый **INITIAL** (начало).

Скелет файла **ЛЕКСИКОН** выглядит следующим образом:

ALTERNATION End End
LEXICON INITIAL
0 End "["
LEXICON End
0 # "]"
END

3.1. Файл морфотактических правил для татарского языка разработан на основе морфотактических схем, включая глагольные и номинативные парадигмы и определяет взаимосвязи между основой и аффиксальными группами. Например, фрагмент морфотактических правил для глагольных парадигм выглядит следующим образом:

ALTERNATION BEGIN VERBSPISOK {VERBSPISOK - список глагольных основ, являющихся начальным входом для анализатора}

Пример.

LEXICON VERBSPISOK
сана verb "V(сана)"
уйла verb "V(уйла)"
…

ALTERNATION verb {*далее идет список аффиксальных классов, которые могут следовать за глаголом*} **REFLEX MODAL NOMINATIVE INFINITIVE PARTICIPAL CONTRARY IMPERATIVE REQUEST CONDITIONAL TENSES CONDJFUTURE1 End** {*указанные аффиксальные классы должны доопределяться далее вплоть до соответствующей группы аффиксов*}

ALTERNATION End End {*признак конца присоединения аффикса или присоединение нулевого аффикса*}

LEXICON INITIAL
0 BEGIN "["
INCLUDE r_verb.lex; {*подключается файл глагольных основ*}
LEXICON REFLEX {группа рефлексивных аффиксов}

В первой части лексикона приводится аффиксальная морфема, далее название класса морфем, которая может следовать за этим аффиксом. Третья составляющая отражает комментарии относительно данного лексического ввода.

+Hn COUSATIVE "+REFLEXIVE(ЫН)"
+HS CONTRARY "+REFLEXIVE(ЫШ)"
LEXICON End
0 # ″]"
END {признак конца файла Лексикон}.

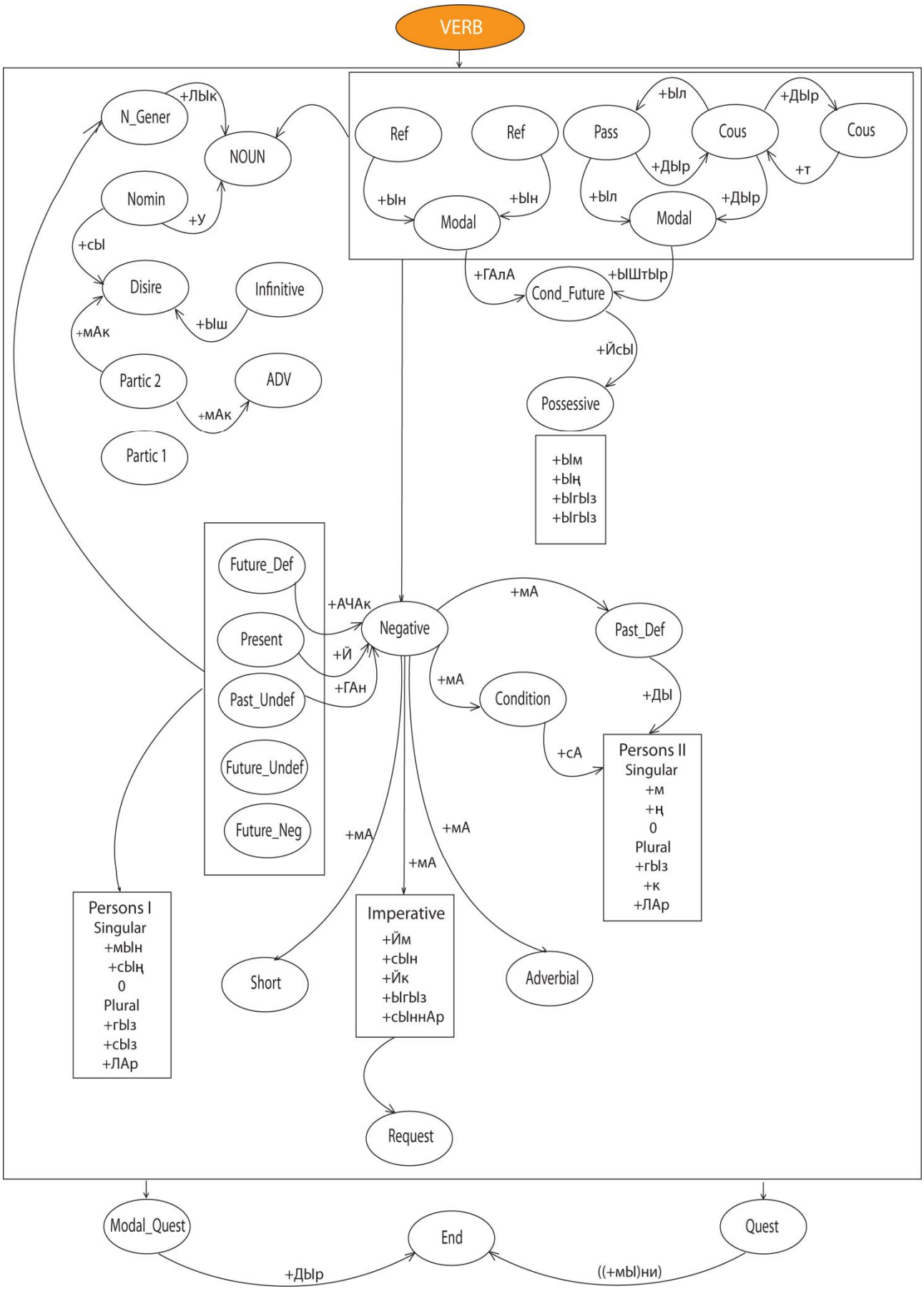

Рис.2. Морфотактическая схема глагольных парадигм.

Морфотактическая схема глагольных парадигм, приведенная на рисунке 2, построена с учетом грамматических категорий наклонения, времени, отрицания, залога, числа и лица глагола. Глагольная основа в словаре представлена в самой краткой форме татарских глаголов, т.е., в форме 2-го лица единственного числа повелительного наклонения: бар - 'иди', кил - 'приходи'. Все аффиксы в схеме приведены в лексическом представлении (ЛП), то есть в зависимости от окружения они обретают разные поверхностные представления (ПП).

Пример.

ЛП: бар (иди)+ГАн кил (приходи)+ ГАн

ПП: барган (сходил) килгән (приходил)

Как видно из примера, здесь аффикс ***-ГАн*** проявляется в двух поверхностных формах: *-ган* и *-гән.*

Лексикон корневых лексем построен на основе современного татарского языка и состоит из 9 лексиконов, заполненных согласно соответствующих требованиям системы: Имена существительные (Nouns), Глаголы (Verbs), Прилагательные (Adjectives), Наречия (Adverbs), Местоимения (Pronouns), Числительные (Numerals), Послелоги (Postpositions), Союзы (Conjunctions), Междометия (Exclamations). Общий объем словаря - 25 900 корневых слов.

Двухуровневый морфологический анализатор построен с использованием грамматики конечных состояний и предназначен для распознавания и генерации словоформ. Рисунок 3 отражает структурно-функциональную схему анализатора.

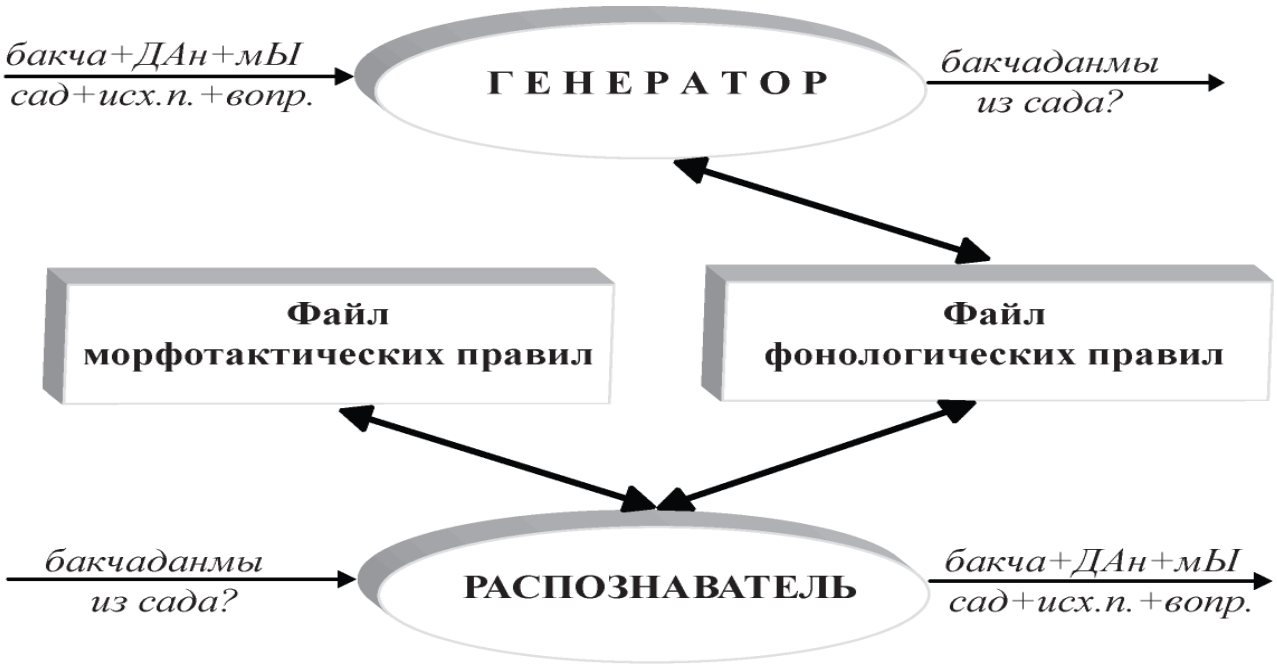

Рис. 3. Структурно-функциональная схема двухуровневого анализатора.

Генератор, используя файл двухуровневых фонологических правил, лексическую запись, например, (*бакча+ДАн+мЫ)* переводит в поверхностную форму – (*бакчаданмы).* Распознаватель, используя оба файла – файл фонологических и файл морфотактических правил, словоформу (поверхностную фор-

му), например, *(бакчадан)* раскладывает по составляющим и соответствующим им описаниям: (**Сущ.**(бакча)+[**падеж.афф.**(*ДАн*)]+[**вопр.афф.**(*мЫ*)]).

В третьей главе на основе анализа систем и методов в области автоматизированных переводчиков сделан вывод о том, что продуктивной и перспективной является концепция и методология программно-концептуальной прагматически-ориентированной технологии для создания переводчиков родственных языков.

Проводится сопоставительный анализ языков на основе объектно-предикативной системы отношений. Описывается методология сопоставления значений аффиксальных морфем на основе объектно-предикативной системы отношений, позволяющая, с одной стороны, эффективно выявлять те или иные различия на глубинном семантическом уровне, с другой стороны, строить лингвистические модели для применения в многоязычных системах обработки данных.

Известно, что значения морфем формируют некий контекст, который наиболее полно раскрывается в семантической ситуации, образуемой словосочетанием, причем каждый аффикс может использоваться в формировании различных контекстов.

Аффиксальные морфемы как минимальные значащие единицы языка, по определению имеют хотя бы одно значение, проявляющееся при использовании его в словоформе. В татарском и турецком языках, зачастую, в зависимости от окружения, аффиксальные морфемы имеют различные интерпретации, т.е. в зависимости от контекста обладают различными значениями, причем одна и та же ситуация не всегда передается одним и тем же классом морфем.

Структура отношений объектно-предикативной системы, используемая для формального представления значений татарских морфем, приведена на рисунке 4.

Для проведения сопоставительного анализа семантики аффиксальных морфем татарского и турецкого языков разработаны специальные фреймовые модели описания объектно-предикативных ситуаций. Это позволяет наиболее полно отразить значения аффиксальных морфем в некотором фрагменте реального мира и строить лингвистические модели перевода, описывающие определенные ситуационные отношения.

Атрибутивные отношения представляют собой ситуации, которые не сочетаются с показателями времени и длительности и называются нединамическими ситуациями[3].

Объектно-ориентированная система, принятая нами за базу сопоставления татарского и турецкого языков, подробно исследуется и описывается в монографии Сулейманова Д.Ш. и Гатиатуллина А.Р.[4]

---

[3] Падучева Е.В. Семантические исследования (Семантика времени и вида в русском языке; Семантика нарратива). – М.: Школа «Языки русской культуры», 1996. – 464 с.

[4] Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель татарских морфем. – Казань: Фэн, 2003. – С. 55-115.

Приведем примеры сопоставительного анализа на ряде атрибутивных отношений, которые представляются наиболее интересными в плане сопоставления рассматриваемых языков.

Рис. 4. Структура отношений объектно-предикативной системы.

1) Пространственные отношения

Общий вид этих отношений выглядит следующим образом:

$$\mathbf{object_1} \rightarrow \mathbf{relation} \rightarrow \mathbf{object_2} \qquad (\mathbf{F_1}),$$

Здесь и далее $\mathbf{F_i}$ обозначает нумерацию абстрактных семантических схем, глубинных формул, относящихся к определенным типам отношений. **<object$_1$>** и **<object$_2$>** – некоторые объекты, причем, **<object$_1$>** находится в некотором отношении **<relation>** к **<object$_2$>**. '→' – означает направленные отношения (связи) между объектами.

Для пространственных отношений введены следующие обозначения:

1. Совпадать в пространстве	equal_local
2. Быть справа	right_of
3. Быть слева	left_of
4. Быть спереди	before_of
5. Быть сзади	after_of
6. Наискосок	Slanting
7. Пересекаться в пространстве	cross
8. Касаться	touch
9. Находиться на	be_on
10. Быть сверху	above
11. Быть снизу	below
12. Находиться в	be_into

Ниже приводится пример сопоставительного анализа для пространственного отношения *be_on* в татарском и турецком языках:

Таблица 1.

Татарский	Турецкий	*‘Стул на столе’*
Urındıq östäldä	*Sandalye masada*	

Ситуационные модели, отражающие соответствующие фразы в таблице 1, имеют следующие представления:

$X_1^{tat} \longrightarrow$ be_on $\longrightarrow X_2^{tat}$ (PS_1^{tat}) $\quad$ $X_1^{tur} \longrightarrow$ be_on $\longrightarrow X_2^{tur}$ (PS_1^{tur})

где X_1^{tat} = *Urındıq*, X_2^{tat} = *östäl* $\quad$ где X_1^{tur} = *Sandalye*, X_2^{tur} = *masa*

Здесь и далее PS_i^{tat} и PS_i^{tur} – это нумерация ситуационных моделей, заполненных конкретными примерами из татарского или турецкого языков соответственно.

Рассмотрим модели перевода для выражения *be_on*.

Модель перевода для выражения пространственного отношения *be_on* может быть представлена следующим образом:

$$PS_1^{tat} <-> PS_1^{tur}$$
$$\text{if}$$
$$[X_1^{tat}] <-> [X_1^{tur}]$$
$$\text{and}$$
$$[X_2^{tat}] <-> [X_2^{tur}],$$
$$\text{где } PS_1^{tat} = [X_1^{tat}\ X_2^{tat} + \text{Case\_Local(DE)}],\ PS_1^{tur} = [X_1^{tur}\ X_2^{tur} + \text{Case\_Local(DA)}]$$

Двусторонняя стрелка ‘**<->**’ означает взаимо-однозначное соответствие составляющих модели.

Модель перевода означает, что ситуационные модели PS_1^{tat} татарского языка и PS_1^{tur} турецкого языка переводимы друг в друга, если выполняются следующие условия:

а) $PS_1^{tat} = [X_1^{tat}\ X_2^{tat} + \text{Case\_Local(DE)}]$, $PS_1^{tur} = [X_1^{tur}\ X_2^{tur} + \text{Case\_Local(DA)}]$, при этом

б) значения абстрактных переменных X_1^{tat} и X_2^{tat} должны являться переводами X_1^{tur} и X_2^{tur}, соответственно.

Как следует их рассмотренных примеров, пространственное отношение *be_on* в татарском и турецком языках задается при помощи присоединения к

аргументу аффиксов *Case_Local(DE)* (локатив1) и *CLocal(DA)* (локатив1), соответственно.

Рассмотрим примеры сопоставительного анализа для пространственного отношения *before_of*.

Таблица 2

Татарский	Турецкий	*‘Стул перед столом’*
Urındıq östäl aldında	*Sandalye masanın önünde*	

Ситуационные модели фраз для выражения *before_of* описываются следующими схемами:

$X_1^{tat} \longrightarrow$ before_of $\longrightarrow X_2^{tat}$ (PS_2^{tat}) $X_1^{tur} \longrightarrow$ before_of $\longrightarrow X_2^{tur}$ (PS_2^{tur})

где X_1^{tat} = *Urındıq*, X_2^{tat} = *östäl* где X_1^{tur} = *Sandalye*, X_2^{tur} = *masa*

Рассмотрим модели перевода для выражения *before_of*.

Для этих примеров пространственное отношение *before_of* может быть представлена следующей моделью перевода:

$$PS_2^{tat} <-> PS_2^{tur}$$

if

$$X_1^{tat} <-> X_1^{tur}$$

and

$$X_2^{tat} <-> X_2^{tur}$$

где $PS_2^{tat} = [X_1^{tat}\ X_2^{tat}$ +(Case_Gen(nHN)|∅) *al*+3POSS_Sing(ZH)+CLocal(DE)],

$PS_2^{tur} = [X_1^{tur}\ X_2^{tur}$ +CGen(nHn) *ön*+P3sg(sH)+CLocal(DA)]

Данная модель перевода означает, что ситуационные модели PS_2^{tat} татарского языка и PS_2^{tur} турецкого языка переводимы друг в друга, если выполняются следующие условия:

а) $PS_2^{tat} = [X_1^{tat}\ X_2^{tat}$ +(Case_Gen(nHN)|∅) *al*+3POSS_Sing(ZH)+CLocal(DE)],

$PS_2^{tur} = [X_1^{tur}\ X_2^{tur}$ +CGen(nHn) *ön*+P3sg(sH)+CLocal(DA)], при этом

б) значения абстрактных переменных X_1^{tat} и X_1^{tat} должны являться переводами X_1^{tur} и X_2^{tur}, соответственно.

Как следует из примеров, отношение *before_of* для татарского языка выражается при помощи послеложной конструкции *‘послелог (al) + аффикс притяжательности + локатив’*. В турецком языке это отношение выражается при помощи аналогичной конструкции, при этом, если в татарском языке зависимый аргумент X_2^{tat} может и не конкретизироваться при помощи аффикса *Case_Gen(nHn)* (генетив), то в турецком языке он всегда конкретизируется.

2) ***Отношения классификации***

Отношения классификации – отношения между двумя простыми или множественными объектами.

Данное отношение выглядит следующим образом:

$$\mathbf{object_1 \longrightarrow class \longrightarrow object_2 \quad (F_2)}$$

Для отношений классификаций введены следующие обозначения:

1. Иметь класс	name_is
2. Класс-подкласс	subclass_of
3. Часть-целое	part_of

4. Элемент-класс	element_of
5. Вышестоящее-нижестоящее	Lower

Рассмотрим соответствующие примеры с отношением *element_of*:

Таблица 3

Татарский	Турецкий	
Studentlar – uquçılarnıñ ber törkeme	*Talipler – öğrencilerin bir grubudur*	*‘Студенты – одна из групп учащихся’*

Фразам из таблицы 3 соответствуют следующие ситуационные модели:

$X_1^{tur} \longrightarrow$ element_of $\longrightarrow X_2^{tur}$ (PS_3^{tur})
где X_1^{tur} = *Talipler*, X_2^{tur} = *öğrenciler*

$X_1^{tat} \longrightarrow$ element_of $\longrightarrow X_2^{tat}$ (PS_3^{tat})
где X_1^{tat} = *Studentlar*, X_2^{tat} = *uquçılar*

Модель перевода для семантических схем PS_3^{tat} и PS_3^{tur} имеет следующий вид:

$$PS_3^{tat} <-> PS_3^{tur}$$
$$\text{if}$$
$$X_1^{tat} <-> X_1^{tur}$$
$$\text{and}$$
$$X_2^{tat} <-> X_2^{tur},$$

где PS_3^{tat} = [X_1^{tat} X_2^{tat} +(Case_Gen(nHN) *ber törkem*+3Poss_Sing(ZH)],
PS_3^{tur} = [X_1^{tur} X_2^{tur} +CGen(nHn) *bir grub*+3Poss_Sing(sH)+DHr]

Модель перевода означает, что ситуационные модели PS_3^{tat} татарского языка и PS_3^{tur} турецкого языка переводимы друг в друга, если выполняются следующие условия:

а) PS_3^{tat} = [X_1^{tat} X_2^{tat} +(Case_Gen(nHN) *ber törkem*+3Poss_Sing(ZH)], а
PS_3^{tur} = [X_1^{tur} X_2^{tur} +CGen(nHn) *bir grub*+3Poss_Sing(sH)+DHr], при этом

б) значения абстрактных переменных X_1^{tat} и X_1^{tat} должны являться переводами X_1^{tur} X_2^{tur}, соответственно.

Для татарского и турецкого языков отношение *element_of* выражается при помощи слов ‘*ber törkem*’ и ‘*bir grup*’, падежных аффиксов CGen(nHN), CGen(nHn) и притяжательных аффиксов 3Poss_Sing(ZH), P3Sg(sH), устанавливающих конкретизирующие отношения между аргументами.

Рассмотрим примеры с отношением *part_of*:

Таблица 4

Татарский	Турецкий	
Malaynıñ qulı	*Erkek çocuğun eli*	*‘Рука мальчика’*

Фразы из таблицы 4 отображаются в следующие ситуационные модели:

$X_2^{tat} \longrightarrow$ part_of $\longrightarrow X_1^{tat}$ (PS_4^{tat})
где X_2^{tat} = *qul*, X_1^{tat} = *Malay*

$X_2^{tur} \longrightarrow$ part_of $\longrightarrow X_1^{tur}$ (PS_4^{tur})
где X_2^{tur} = el, X_1^{tur} = *Erkek çocuk*

$$PS_4^{tat} <-> PS_4^{tur}$$
$$\text{if}$$
$$X_1^{tat} <-> X_1^{tur}$$
$$\text{and}$$
$$X_2^{tat} <-> X_2^{tur},$$

где PS_4^{tat} = [X_1^{tat} +Case_Gen(nHN) X_2^{tat} +3Poss_Sing(ZH)],
PS_4^{tur} = [X_1^{tur} +CGen(nHn) + X_2^{tur} +P3sg(sH)]

Ситуационные модели PS_4^{tat} и PS_4^{tur} переводимы друг в друга, если выполняются следующие условия:

а) $PS_4^{tat} = [X_1^{tat} + Case\_Gen(nHN)\ X_2^{tat} + 3Poss\_Sing(ZH)]$, а

$PS_4^{tur} = [X_1^{tur} + CGen(nHn) + X_2^{tur} + P3sg(sH)]$, при этом

б) значения абстрактных переменных X_1^{tat} и X_1^{tat} должны являться переводами X_1^{tur} и X_2^{tur}, соответственно.

Отношение *part_of* для обоих языков выражается при помощи одинакового типа падежных и притяжательных аффиксов Case_Gen(nHN), P3sg(ZH) и CG en(nHn), P3sg(sH), соответственно, в татарском и турецком языках.

3) ***Отношения принадлежности***

Отношение принадлежности является отношением между двумя объектами и имеет следующий вид:

$$\mathbf{object_1 \longrightarrow belong \longrightarrow object_2} \qquad (F_3)$$

Рассмотрим пример сопоставительного анализа для отношения принадлежности *belong*.

Таблица 5

Татарский	Турецкий	*'Мой стул'*
Minem urındıq	*Sandalyem*	

Ситуационные модели для выражений из таблицы 5 с отношением *belong* имеют следующие представления:

$X_2^{tat} \longrightarrow$ belong $\longrightarrow X_1^{tat}$, (PS_5^{tat}) где $X_2^{tat} =$ *urındıq*, $X_1^{tat} =$ *Min+em*

$X_2^{tur} \longrightarrow$ part_of $\longrightarrow X_1^{tur}$, (PS_5^{tur}) где $X_2^{tur} =$ *Sandalye+m*, $X_1^{tur} = \varnothing$

Таким образом, модель перевода, выражающая отношение принадлежности для семантических схем PS_5^{tat} и PS_5^{tur}, будет иметь следующий вид:

$$PS_5^{tat} <-> PS_5^{tur}$$
$$X_1^{tat} = min$$
$$X_1^{tur} = \varnothing$$
$$X_2^{tat} <-> X_2^{tur},$$

где $PS_5^{tat} = [X_1^{tat} + P1sg(Hm)\ X_2^{tat}]$, $PS_5^{tur} = [X_2^{tur} + P1sg(Hm)]$ или

$PS_5^{tat} = [X_1^{tat} + P1sg(Hm)\ X_2^{tat} + P1sg(Hm)]$, $PS_5^{tur} = [X_2^{tur} + P1sg(Hm)]$

Таким образом, ситуационные PS_5^{tat} и PS_5^{tur} переводимы друг в друга, если выполняются следующие условия:

а) $PS_5^{tat} = [X_1^{tat} + P1sg(Hm)\ X_2^{tat}]$, а

$PS_5^{tur} = [X_2^{tur} + P1sg(Hm)]$ или

$PS_5^{tat} = [X_1^{tat} + P1sg(Hm)\ X_2^{tat} + P1sg(Hm)]$, а

$PS_5^{tur} = [X_2^{tur} + P1sg(Hm)]$, при этом

б) значение абстрактной переменной X_2^{tat} должно являться переводом X_2^{tur}, а $X_1^{tat} = min$ и $X_1^{tur} = \varnothing$ (пустое значение).

В отличие от татарского языка, в турецком языке при выражении отношения принадлежности между двумя объектами, если объект принадлежности является личным местоимением, само личное местоимение явно не присутствует и выражается лишь наличием аффикса притяжательности. Отличи-

тельной особенностью турецкого языка является также и то, что группа слов, обозначающая названия родственных отношений, в любом контексте употребляется с аффиксом принадлежности.

В диссертации описывается метод и реализуется алгоритм машинного обучения (МО) для перевода в тюркских языках на основе механизмов сопоставления лексических уровней соответствий между языками. Обучающая выборка формируется из множества переведенных пар предложений.

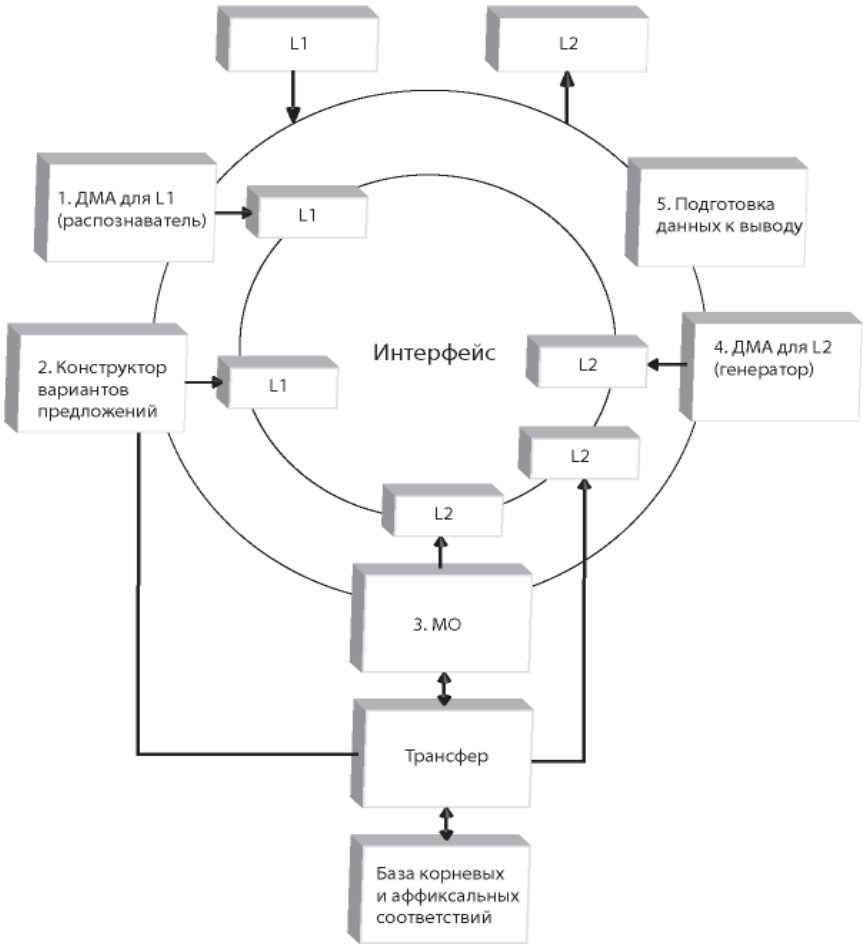

Рис. 5. Структурно-функциональная схема для татарско-турецкого машинного переводчика.

Алгоритм МО предполагает существование моделей перевода, использующих сходные и различные части предложений между двумя переведенными парами (e_i, e_j) из двух параллельных блоков. Формально, переведенный пример составлен из пары предложений, которые являются переводами друг друга из языка **L1** в язык **L2**, соответственно.

В двух данных переведенных парах (e_i, e_j) мы пытаемся найти сходства между составляющими e_i и e_j. Предложения рассматриваются как последовательность лексических знаков (т.е. слов или морфем). Если не найдено никаких сходств, то на этих примерах модели не обучаются. Если есть сходные составляющие, то генерируется парная последовательность в следующей форме:

$s^1_0, d^1_0, s^1_1, \ldots, d^1_{n-1}, s^1_n, <\!\!-\!\!> s^2_0, d^2_0, s^2_1, \ldots, d^2_{m-1}, s^2_m,$ где n, m >= 1

s^1_k представляет сходства (последовательность общих знаков) между e^1_i и e^1_j. Подобным образом, d^1_k: $(d^1_{i,k}, d^1_{j,k})$ представляют различия между e^1_i и e^1_j, где

$\mathbf{d^1_{i,k}}$, $\mathbf{d^1_{j,k}}$ непустые различные знаки между двумя сходными элементами $\mathbf{s^1_k}$ и $\mathbf{s^1_{k+1}}$. Соответствующие различия не содержат общих знаков. Т.е., для $\mathbf{d_k}$ различия, $\mathbf{d_{i,k}}$, $\mathbf{d_{j,k}}$ не содержат никаких общих знаков. Также, ни один общий знак сходности $\mathbf{s_i}$ не появляется ни в одном ранее образованном различии $\mathbf{d_k}$, для всех $\mathbf{k < i}$. Любые из $\mathbf{s^1_0}$, $\mathbf{s^1_n}$, $\mathbf{s^2_0}$ или $\mathbf{s^2_m}$ могут быть пустыми, но для любого $\mathbf{0 < i < n}$ и $\mathbf{0 < j < m}$, $\mathbf{s^1_i}$ и $\mathbf{s^2_j}$ не должны быть пустыми. Заметим, что между двумя образцами переведенных пар существует либо одно соответствие, либо ни одного.

На основе разработанных математических лингвистических моделей реализованы программные модули системы машинного перевода для тюркских языков, в частности, программные модули системы татарско-турецкого машинного перевода. На рисунке 5 приведена структурно-функциональная схема данного программного комплекса. Модульная структура программного комплекса содержит пользовательскую и алгоритмические части, при этом алгоритмическая часть является языконезависимой, что при необходимости позволяет строить модели перевода для разных языков.

Рассмотрим этапы обработки текста на примере татарско-турецкого перевода. Пусть на вход системы поступает следующая последовательность словоформ, образующая следующее предложение на татарском языке *«Мин көзге юлдан бардым» ‘Я ходил по осенней дороге’*. Ниже приведены результаты обработки данной последовательности в порядке выполнения модулей, указанном в структурно-функциональной схеме татарско-турецкого перевода:

1) Двухуровневый морфологический анализатор (ДМА) с функцией распознавания, используя файлы морфотактики и фонологических правил, выдает проанализированные словоформы с приписанными морфологическими характеристиками:

1. мин	[Pro1_Sing(мин)]	*‘Мест.(я)’*
2.1. көзге	[N(көз) +CASE_POINT(ГЫ)]	*‘Сущ.(осень)+Пад.афф.принадл.(ГЫ)’*
2.2. көзге	[N(көзге)]	*‘Сущ.(зеркало)’*
3. юлдан	[N(юл)+CASE_ABL(ДАн)]	*‘Сущ.(дорога)+Исх.пад.(ДАн)’*
4.1. бардым	[V(бар)+POST_DAF(ДЫ)+1PS_Sing(м)]	*‘Гл.(иди)+Прош.вр.(ДЫ)+1л.ед.ч.(м)’*
4.2. бардым	[N(бард)+1POSS_Sing(ЫМ)]	*‘Сущ.(бард)+Прит. 1л.ед.ч.(ЫМ)’*

2) Результат морфологического разбора словоформы, как видно из примера, имеет большое число лексических неопределенностей. Конструктор вариантов предложений формирует всевозможные варианты предложений:

а) [Pro1_Sing(мин)] [N(көз)+CASE_POINT(ГЫ)] [N(юл)+CASE_ABL(ДАн)] [V(бар)+POST_DAF(ДЫ)+1PS_Sing(м)]

б) [Pro1_Sing(мин)] [N(көз)+CASE_POINT(ГЫ)] [N(юл)+CASE_ABL(ДАн)] [N(бард)+1POSS_Sing(ЫМ)]

в) [Pro1_Sing(мин)] [N(көзге)] [N(юл)+CASE_ABL(ДАн)] [V(бар)+POST_DAF(ДЫ)+1PS_Sing(м)]

г) [Pro1_Sing(мин)] [N(көзге)] [N(юл)+CASE_ABL(ДАн)] [N(бард)+1POSS_Sing(ЫМ)]

3) Все варианты предложений поступают на вход подсистемы МО, где осуществляется перевод путем выбора наиболее соответствующих моделей перевода.

Для предложенных вариантов подсистема МО выдаст единственную модель перевода в виде:
Мин X_1^{tat}+CASE_POINT(ГЫ) X_2^{tat}+CASE_ABL(ДАн) X_1^{tat}+POST_DEF(ДЫ)+1PS_Sing(м) <=> Ben X_1^{tur} X_2^{tur} +P3sg+Cabl X_3^{tur}+ TAM1past(DH)+P1_sing(m) If Мин=Ben, X_1^{tat} =X_1^{tur} and X_2^{tat} =X_2^{tur} and X_3^{tat} =X_3^{tur};

Далее происходит замена элементов модели перевода на базе аффиксальных и корневых соответствий: *ben sonbahar yol+sH+DAn yürü+DH+m*

4) Двухуровневый морфологический анализатор (ДМА) с функцией генерации, используя двухуровневые правила для турецкого языка выдаст следующую сгенерированную последовательность: *ben sonbahar yolından yürüdüm*

5) Модуль подготовки выходных данных позволяет выводить данные с соответствующим форматированием.

Программный комплекс реализован для операционной системы не ниже Windows'98 и представляет собой единый исполняемый модуль в объеме 680,5 КБ. Морфотактическая база для татарского языка занимает 1664 КБ. Количество двухуровневых автоматных правил для татарского языка составляет 47 правил. Количество моделей перевода, полученных в результате выполнения алгоритма машинного обучения составляет 138 моделей.

Заключение

Диссертационная работа посвящена проблеме создания математических лингвистических моделей и их эффективной реализации. В процессе выполнения работы получены следующие результаты:

1. Разработана полная компьютерная модель татарской морфологии в виде двухуровневой автоматной модели.
2. Разработан программный инструментарий для морфологического анализа и синтеза татарских текстов на основе двухуровневой автоматной модели.
3. Разработаны формальные семантические модели аффиксальных морфем на основе объектно-предикативных схем и проведен сопоставительный анализ семантических схем для тюркских языков.
4. Разработаны формальные модели перевода на основе алгоритмов машинного обучения, использующие шаблоны переводных соответствий языков.
5. Реализованы программные модули в составе системы татарско-турецкого машинного перевода.

В Приложении 1 содержатся акты о внедрениях и справки об использовании программного комплекса, разработанного и реализованного в рамках данной диссертационной работы.

В Приложении 2 приводится полный файл двухуровневых автоматных правил.

В Приложении 3 приводится пример генерации словоформы с падежным аффиксом -*ЛАр* на базе описанных фонологических правил.

В Приложении 4 приводится полное описание файла морфотактических правил.

В Приложении 5 приводится демонстрационный пример выполнения функции распознавания для поверхностной формы: *уйнарга (‘играть’)*.

В Приложении 6 приводятся лингвистические модели, полученные в результате выполнения алгоритма МО.

Список публикаций по теме диссертации

Публикации в рецензируемых журналах, рекомендованных ВАК:

1. Гильмуллин Р.А. Модуль обучающейся модели татарско-турецкого машинного переводчика // Вестник Казанского государственного технического университета им. А.Н.Туполева. - 2007, № 2(46) - С. 65-67.

Прочие публикации

2. Гильмуллин Р.А. Реализация контекстных соответствий Ы:ы, Ы:е и Ы:0 в файле фонологических правил // Сборник трудов Математического центра имени Н.И. Лобачевского. Т.4. Компьютерная лингвистика. – Казань: УНИПРЕСС, 1999. – С. 51-58.

3. Гильмуллин Р.А. К разработке татарско-турецкого машинного переводчика // Труды Казанской школы-семинара по компьютерной и когнитивной лингвистике TEL-2001. Выпуск 6. – Казань: Из-во “Отечество”, 2001, – С. 12-18.

4. Гильмуллин Р.А. Разработка файла морфотактических правил для глагольных групп татарского языка // Проблемы сохранения языка и культуры в условиях глобализации: Материалы VII Международного Симпозиума "Языковые контакты Поволжья" / Науч.ред. И.А.Гилязов. – Казань: КГУ, 2009. – С. 222-226.

5. Suleymanov D.Sh., Guilmullin R.A., Guilmy A.A. Two-level phonological rules of Tatar morphology // Научные труды YI международной конференции "Знания-Диалог-Решение". – Крым, Ялта. 15-20 сентября 1997. – С. 299-305. (в соавторстве, 30% личного участия)

6. Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. Двухуровневое описание морфологии татарского языка // Тезисы Международной научной конференции, посвященной 200-летию университета: "Языковая семантика и образ мира". 7-10 октября 1997. Книга 2. – Казань: Изд-во КГУ. – С. 65-67. (в соавторстве, 30% личного участия)

7. Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. Файл фонологических правил татарского языка // Электронная конференция информационные технологии в гуманитарных науках 25-31 мая, 1998. – Казань: HTTP://www.kcn.ru/_tat_ru/universitet/gum_konf/ot7.htm. (в соавторстве, 50% личного участия)

8. Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. База морфотактических правил для татарского глагола как основа двухуровневого морфоло-

гического анализатора // Сборник трудов Международного семинара ДИАЛОГ-98. – Казань, 1-2 июня. – С. 597-609. (в соавторстве, 50% личного участия)

9. Сулейманов Д.Ш., Гильмуллин Р.А. Реализация контекстных соответствий A:a, A: ä в файле фонологических правил // Сборник трудов Математического центра имени Н.И. Лобачевского. Т.4. Компьютерная лингвистика. – Казань: УНИПРЕСС, 1999. – С. 127-137. (в соавторстве, 50% личного участия)

10. Сулейманов Д.Ш., Гильмуллин Р.А. Реализация контекстных соответствий V:u, V: ü, V:0, Y:I и Y: ö в файле фонологических правил // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2000. Выпуск 5. – Казань: Из-во “Сэлэт”, 2001, – С. 162-167. (в соавторстве, 50% личного участия)

11. Сулейманов Д.Ш., Гильмуллин Р.А. Реализация контекстных соответствий Д:н, Д: д, Д:т, Л:н, С:с в файле фонологических правил // Сборник трудов Международного семинара ДИАЛОГ-2000: Компьютерная лингвистика и её приложения. Т. 2. Прикладные проблемы. – Протвино, июнь. – С. 396-404. (в соавторстве, 50% личного участия)

12. Сулейманов Д.Ш., Невзорова О.А., Салимов Ф.И., Гильмуллин Р.А. Автоматизированный перевод документов в системах учета и регистрации: концептуально-алгоритмическая модель // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2001. – Казань: Из-во “Отечество”, 2001, – С. 108-120. (в соавторстве, 30% личного участия)

13. Гильмуллин Р.А., Ишимов В.В. К разработке татарско-турецкого машинного переводчика // Компьютерная лингвистика и интеллектуальные технологии. Тр. Междунар. семинара Диалог’2002. Т.2.: Прикладные проблемы. – М.: Наука, 2002. – С. 133-138. http://dialog-21.ru/archive_article.asp?param=7544&y=2002&vol=6078. (в соавторстве, 70% личного участия)

14. **Suleymanov** D.Sh.**, Guilmoulline R.A., Guilmoulline A.A. Tatar phonological rules as a base of two-level morphological analyzer, in Proceedings of LP’2000, ed. B.Palek and O.Fujimura: 495-504 p., The Karolinum Press, Prague**. (в соавторстве, 30% личного участия)

15. Гильмуллин Р.А., Минабова Э.К. Сопоставительный анализ семантики аффиксальных морфем в татарском и турецком языках на основе объектно-предикативной системы отношений // Международный симпозиум «Типология аргументной структуры и синтаксических отношений» Тезисы докладов. Казань, 2004. – С. 323-236. (в соавторстве, 70% личного участия)

16. Сулейманов Д.Ш., Гильмуллин Р.А., Сафина Л.Р. Использование компьютерных технологий в обучении: на примере обучающе-тестирующей программы «Морфологический анализатор» // Международный журнал «Образовательные технологии и общество», том 9, № 4, 2006. – Казань: http://ifets.ieee.org/russian/depository/v9_i4/pdf/7.pdf (в соавторстве, 30% личного участия)

17. Сулейманов Д.Ш., Невзорова О.А., Гатиатуллин А.Р., Гильмуллин Р.А., Аюпов М.М., Пяткин Н.В. Основные компоненты прикладной грамматической модели татарского языка // Компьютерная лингвистика и интеллекту-

альные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая - 3 июня 2007 г.) / Под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. - М.: Изд-во РГГУ, 2007. 658 с.: ил. – С. 525-530. (в соавторстве, 25 % личного участия)

18. Сулейманов Д.Ш., Хакимов Б.Э., Гильмуллин Р.А. Из опыта татарской локализации ОС Windows и офисных приложений // Проблемы сохранения языка и культуры в условиях глобализации: Материалы VII Международного Симпозиума "Языковые контакты Поволжья" / Науч.ред. И.А.Гилязов. – Казань: КГУ, 2009. – С. 226-230. (в соавторстве, 30% личного участия)

19. Хакимов Б.Э., Гильмуллин Р.А. К разработке системы параметров морфологической разметки для электронного корпуса татарских текстов // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2008. – Казань: Казан. гос. ун-т, 2009. – С. 24-29. (в соавторстве, 50% личного участия)

20. D.Sh. Suleymanov, R.A. Gilmullin Comparative Analysis of Meanings of Affixal Morphemes in the Tatar and Turkish Languages for Machine Translation Tasks // Interactive Systems and Technologies: the Problems of Human-Computer Interaction. Volume III. – Collection of scientific papers. – Ulyanovsk: UlSTU, 2009. – 312-320 p. (в соавторстве, 70 % личного участия)